Centrum voor Wiskunde en Informatica
REPORT *RAPPORT*

Inference on Rare Errors Using Asymptotic Expansions and
Bootstrap Calibration

R. Helmers

Probability, Networks and Algorithms (PNA)

PNA-R9820 December 1998

Report PNA-R9820
ISSN 1386-3711

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Inference on Rare Errors Using Asymptotic Expansions and Bootstrap Calibration

R. Helmers
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT
The number of items in error in an audit population is usually quite small, whereas the error distribution is typically highly skewed to the right. For applications in statistical auditing, where line item sampling is appropriate, a new upper confidence limit for the total error amount in an audit population is obtained. Our method involves an empirical Cornish-Fisher expansion in the first place; in the second stage we employ the bootstrap to calibrate the coverage probability of the resulting interval estimate.

1991 Mathematics Subject Classification: 62E17, 62G09, 62G15, 62E25.
Keywords and Phrases: Auditing, bootstrap calibration, Cornish-Fisher expansion, finite population, line item sampling, non-standard mixtures, Poisson distribution, rare errors, superpopulation model.
Note: The work is carried out in PNA 3.2 Statistics.

1. Introduction

The problem investigated in this paper arises in statistical auditing. Consider a finite population of N items with recorded values $y_1, \dots, y_N$ the 'book amounts'. Suppose that the items may be subject to (unknown) errors $e_1, \dots, e_N$, and that $x_1, \dots, x_N$ are the 'true' values, the 'audited' amounts of the N items. Thus, $e_i = y_i - x_i$ denotes the 'error amount' corresponding to the book value y_i of the ith item (audit unit), $i = 1, \dots, N$. However, it is known a priori that most of the e_i's are zero, and the auditor's problem is to give a $(1-\alpha)$ upper confidence bound for the total population error amount

$$D = \sum_{i=1}^{N} e_i \tag{1.1}$$

when a random sample S of book amounts of size n, drawn without replacement, from the population $\{y_1, \dots, y_N\}$ is available, and $e_i,\ i \in S$ denote the errors observed by the auditor in the recorded values $Y_i,\ i \in S$. Clearly $\sum_{i \in S} e_i$ is the total error amount in the sample, and $\hat{D}_n = \frac{N}{n} \sum_{i \in S} e_i$ is an unbiased estimator of D, our parameter of interest.

Let p denote the (small) probability that e_i is non-zero, i.e. that item i, with true value x_i, is in error, and let M be the number of items in the sample $Y_1, \dots, Y_n$ with error. Clearly

$$\hat{D}_n = \frac{N}{n} \sum_{j=1}^{M} V_j \tag{1.2}$$

where the V_j's are the observed non-zero error amounts in the sample. In typical applications, errors are rare, i.e. p is close to zero, and the sample size n is small compared with the size of the population N. In such cases one may impose a superpopulation model on $\{y_1, \dots, y_N\}$, where $y_i = x_i + e_i,\ i = 1, \dots, N$, by assuming that the e_i's are independent random variables with a common distribution of non-standard mixture type (cf Tamura (1989))

$$F = pG + (1-p)\delta_0 \tag{1.3}$$

where δ_0 denotes the degenerate distribution which puts all its probability mass at the point zero. Clearly the V_j's constitute a random sample of random size M from the unknown non-zero error amount distribution G, while M is the (random) number of non-zero errors present in a random sample of size n from F. In such cases one may assume that M is Poisson(ν)-distributed, with unknown parameter $\nu = np$, in addition M is assumed to be independent of the V_j's. The Poisson approximation for M works well, provided the error rate p is small. There is no use for the classical requirement that $\nu = np$ is fixed, while $p \downarrow 0$ and $n \to \infty$. On the contrary, in the present paper we let $\nu = \nu_n$ approach infinity, as $n \to \infty$, whereas p is assumed to be small but fixed. We refer to Barbour et.al. (1992) for an excellent account of the theory of Poisson approximations.

Let $\mu = \int x dG(x)$ denote the mean of G; G is nothing but the distribution of V_1, i.e. the conditional distribution of an error amount e, given that $e \neq 0$. Because $n << N$ one may argue that for our purposes D - the total population error in the finite population under consideration - can be replaced by $E_F D = \frac{N}{n}\nu\mu$, under random sampling from the superpopulation error distribution F. (The only exception would be the case that ν is extremely small, but we rule out this case here). The problem now becomes to find a $(1-\alpha)$ upper confidence limit for $E_F D = \frac{N}{n}\nu\mu$. Note that G is typically highly skewed to the right.

In statistical auditing items are often selected without replacement with probability proportional to recorded book values (e.g. by applying dollar-unit sampling). In the present paper, however, we employ simple random sampling without replacement (i.e. audit-unit or line-item sampling). This appears more convenient in a variety of situations where to ascertain the correct value for each audit unit, i.e. each recorded value y_i, is equally important, and should have equal chance to be included in our sample. E.g. in social security, where payments of disability or unemployment benefits should be correct, irrespectively whether the benefit is a large or a small amount. Also in tax examinations and other audit applications in the public sector the auditor employs line item of audit-unit sampling (cf. Tamura (1989), p.6).

In this paper we establish an upper confidence limit with confidence level at least equal to $(1-\alpha)$ for the total population error D using asymptotic expansions and bootstrap calibration. Our focus is on the important situation that errors are rare and the non-zero error distribution is highly skewed. In realistic cases G may consist of a finite mixture of light-tailed distributions (like the exponential). It is well-known that such mixtures are hard to distinguish from heavy tail models (cf. Jensen (1995), Chapter 7). Hence one should not only correct for skewness but for kurtosis as well. Our method will give a much better one-sided confidence interval for D than the traditional normal approximation can provide us with. However, a cautionary remark appears to be in order here: *no* method for setting confidence limits for D will work in *all* cases. For example, I would not dare to apply the method proposed in this paper for cases, where $M = 0$ or 1 and sample sizes as small as $n = 100$, say.

2. Asymptotic expansions

As the normal approximation behaves typically rather poorly in audit populations one may try - as a first step - to improve upon this by employing Cornish-Fisher expansions. The idea is to adapt for skewness and kurtosis by estimating the third and fourth cumulants appearing the Cornish-Fisher expansion from the data set of observed non-zero error amounts at hand. However, one cannot really expect that the empirical Cornish-Fisher expansion will work well in most instances, as our estimates of the third and fourth cumulant are by necessity highly variable, because the number of non-zero errors is usually quite small. To deal with this shortcoming we employ bootstrap calibration (cf section 3) which presumably will extend the range of validity of our method considerably.

To begin with the analysis, let us first note that we will assume throughout that p is fixed, but close to zero (so that the Poisson approximation for M is applicable), while at the same time the sample size n approaches infinity ($n \to \infty$). This, of course, directly yields that the expected number of non-zero errors in the sample $EM = \nu = np$ - though only a small fraction of the sample size n - gets large as well in the asymptotics. A simple calculation gives

$$E_F\hat{D}_n = \nu\mu\frac{N}{n},\ \sigma_F^2(\hat{D}_n) = \nu\mu_2\frac{N^2}{n^2} \tag{2.1}$$

where $\mu = E_G V_1,\ \mu_2 = E_G V_1^2$. The third and fourth cumulant κ_{3n} and κ_{4n} of $\hat{D}_n$ are also easily found:

$$\begin{aligned} &\kappa_{3n} = E_F(\hat{D}_n - E\hat{D}_n)^3/\sigma_F^3(\hat{D}_n) \\ \sim\quad &\kappa_3 = \frac{\tilde{\mu}_3 + 3\sigma^2\mu + \mu^3}{\nu^{1/2}\mu_2^{3/2}} \end{aligned} \tag{2.2}$$

with $\sigma^2 = \sigma_G^2(V_1)$, $\tilde{\mu}_3 = E_G(V_1-\mu)^3$, and

$$\begin{aligned} &\kappa_{4n} = E_F(\hat{D}_n - E\hat{D}_n)^4/\sigma^4(\hat{D}_n)\ -\ 3 \\ \sim\quad &\kappa_4 = \frac{\tilde{\mu}_4 + 4\mu\tilde{\mu}_3 + 6\sigma^2\mu^2 + \mu^4}{\nu\mu_2^2} \end{aligned} \tag{2.3}$$

where $\tilde{\mu}_4 = E_G(V_1-\mu)^4$. The error committed in the approximations (2.2) and (2.3) is of order $(\frac{p}{n})^{1/2}$ and $\frac{p}{n}$ respectively; $\sim$ refers to the fact that we have deleted such errors. The quantities κ_3 and κ_4 are easily checked to be exactly equal to the third and fourth cumulant of $\sum_{j=1}^M V_j$, where the V_j's denote a random sample from G, with Poisson(ν) distributed random sample size M.
Define Studentized statistics $S_{1,n}$ and $S_{2,n}$ by

$$S_{1,n} = \frac{\hat{D}_n - \nu\mu\frac{N}{n}}{(\sum_{j=1}^M V_j^2)^{1/2}\frac{N}{n}},\ S_{2,n} = \frac{\hat{D}_n - \nu\mu\frac{N}{n}}{Nn^{-1/2}\hat{s}} \tag{2.4}$$

where $\hat{s}^2 = n^{-1}\sum_{j=1}^n (e_j - \bar{e})^2$, with $\bar{e} = n^{-1}\sum_{j=1}^n e_j$.
Note that

$$P_F(S_{1,n} \le x) = P_F(S_{2,n}Q^{\frac{1}{2}} \le x) \tag{2.5}$$

where

$$Q = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{\sum_{j=1}^M V_j^2} \tag{2.6}$$

A simple computation yields

$$Q = 1 - \frac{2M^2\bar{V}^2}{n\sum_{j=1}^M V_j^2} + R \tag{2.7}$$

while it is easily seen that R is a non-negative random term at most equal to $\frac{M^2}{n^2}$. Note that R is of negligible order for our purposes, when p is close to zero and n gets large. The second term on the right of (2.7) is easily seen to be of order p (as $p \downarrow 0$) in probability. Hence, it is easily verified that

$$P_F(S_{1,n} \le x) \ge P_F(S_{2,n} \le x) + o(p) \tag{2.8}$$

for $x \ge 0$, while the reverse inequality holds for $x < 0$. Here the o-term in (2.8) is a remainder term of lower order, uniformly in x, negligible for our purposes when p is close to zero and n gets large.

The distribution of $S_{2,n}$ is the distribution of the classical Student t-statistic $n^{\frac{1}{2}}(\bar{e} - p\mu)/\hat{s}$, based on a sample of size n from F. Of course $\int x\ dF(x) = p\mu$. Let $c_{2,n,\alpha}$ denote the $(1-\alpha)$th critical point of $S_{2,n}$, i.e. $P(S_{2,n} \le c_{2,n,\alpha}) = 1-\alpha$. In Hall(1988), pp.944-945, example 1, one can find a Cornish-Fisher expansion of $c_{2,n,\alpha}$:

$$\begin{aligned} c_{2,n,\alpha} \quad \sim \quad & u_\alpha + \frac{(2u_\alpha^2+1)}{6}\kappa_{3n} \\ & + u_\alpha\{-\frac{1}{12}\kappa_{4n}(u_\alpha^2-3) + \frac{5}{72}\kappa_{3n}^2(4u_\alpha^2-1) + \frac{1}{4}n^{-1}(u_\alpha^2+3)\} \end{aligned} \tag{2.9}$$

where κ_{3n} and κ_{4n} are the third and fourth cumulants of $\bar{e} = n^{-1}\sum_{i=1}^n e_i$ (i.e. of $\hat{D}_n$) under random sampling from F; $\sim$ refers here to the fact that we have deleted terms of smaller order than n^{-1}. The expansion (2.9) has a remainder of order $o(n^{-1})$ as n gets large, provided F possesses an absolutely continuous component and a fourth moment of F exists. (Hall, 1987). Note that p is close to zero, but assumed to be fixed; i.e. F is also fixed in the asymptotics, as required by Hall (1987; 1988); otherwise the non-singularity requirement may cause problems.
Let now $c_{1,n,\alpha}$ denote the $(1-\alpha)$th critical point of $S_{1,n}$, i.e. $P(S_{1,n} \le c_{1,n,\alpha}) = 1-\alpha$, where (cf. (2.4)) $S_{1,n} = (\sum_{j=1}^M V_j - \nu\mu)/(\sum_{j=1}^M V_j^2)^{1/2}$. If $0 < \alpha < \frac{1}{2}$, then, obviously $c_{1,n,\alpha} \le c_{2,n,\alpha}$ for sufficiently large n. Hence, Hall's expansion (2.9) for $c_{2,n,\alpha}$ provides us with a *conservative* approximation for $c_{1,n,\alpha}$ for confidence levels $1-\alpha$ with $0 < \alpha \le \frac{1}{2}$ for sufficiently large n, provided p is close to zero. It follows that an approximate upper confidence limit with confidence level at least equal to $(1-\alpha)$ for $E_F D = \frac{N}{n}\nu\mu$ is

$$\hat{D}_n + c_{2,n,\alpha}\frac{N}{n}(\sum_{j=1}^M V_j^2)^{\frac{1}{2}} \tag{2.10}$$

provided the error rate p is small enough and $0 < \alpha < \frac{1}{2}$. However, the upper bound (2.10) cannot be computed from the data, as the 'theoretical critical point' $c_{2,n,\alpha}$ (Hall (1988)) is unknown, because G is unknown. Hence we replace (2.10) by its empirical counterpart,

$$\hat{D}_n + \hat{c}_{2,n,\alpha}\frac{N}{n}(\sum_{j=1}^M V_j^2)^{\frac{1}{2}} \tag{2.11}$$

where $\hat{c}_{2,n,\alpha}$ is approximated by

$$\begin{aligned} \hat{c}_{2,n,\alpha} \quad \sim \quad & u_\alpha + \frac{(2u_\alpha^2+1)}{6}\hat{\kappa}_3 \\ & + u_\alpha\{-\frac{1}{12}\hat{\kappa}_4(u_\alpha^2-3) + \frac{5}{72}\hat{\kappa}_3^2(4u_\alpha^2-1) + \frac{1}{4}M^{-1}(u_\alpha^2+3)\}, \end{aligned} \tag{2.12}$$

with

$$\hat{\kappa}_3 \quad = \quad \frac{\hat{\hat{\mu}}_3 + 3\hat{\hat{\mu}}_2\bar{V} + \bar{V}^3}{M^{-1}(\sum_{j=1}^M V_j^2)^{\frac{3}{2}}}$$

and

$$\hat{\kappa}_4 \quad = \quad \frac{\hat{\hat{\mu}}_4 + 4\hat{\hat{\mu}}_3\bar{V} + 6\hat{\hat{\mu}}_2\bar{V}^2 + \bar{V}^4}{M^{-1}(\sum_{j=1}^M V_j^2)^2}. \tag{2.13}$$

where $\hat{\hat{\mu}}_l = M^{-1}\sum_{j=1}^M (V_j - \bar{V})^l$, for $l = 2,3,4$ and $\bar{V} = M^{-1}\sum_{j=1}^M V_j$. Clearly, the coverage probability of empirical Cornish-Fisher bound (2.11) satisfies the inequality

$$\lim_{n\to\infty} P_F(E_F D < (2.11)) \ge 1-\alpha, \tag{2.14}$$

provided the error rate p is small enough and $0 < \alpha < \frac{1}{2}$. In (2.12) $\sim$ indicates that, in addition to the error terms already deleted in the previous steps, the random approximation error in (2.12) (due to replacing our Cornish-Fisher expansion by its empirical counterpart) is of smaller order in probability than ν^{-1}, as $\nu \to \infty$.

3. Bootstrap Calibration

The empirical Cornish-Fisher bound (2.11) is easy to compute. However, the coverage probability (cf. (2.14)) $P_F(E_F D < (2.11))$ may in fact not be at least equal to the nominal confidence level $1-\alpha$, as desired in finite samples. To remedy this defect one may employ bootstrap calibration (cf. Beran (1987), Hall and Martin (1988)). The idea is to estimate by means of resampling the coverage probability, with α replaced by λ, for a grid of values of λ in (0,1), and select the largest value $\hat{\lambda}$ for which the bootstrap estimate

$$P_n^*(\hat{D}_n < \hat{D}_n^* + \hat{c}^*_{2,n,\lambda}\frac{N}{n}(\sum_{j=1}^{M^*} V_j^{*2})^{\frac{1}{2}}) \tag{3.1}$$

is at least $1-\alpha$. Here P_n^* refers to probability in our 'bootstrap world': conditionally given $(V_1, \ldots, V_M)$, a bootstrap resample $(V_1^*, \ldots, V_{M^*}^*)$ of size M^* is drawn with replacement from $(V_1, \ldots, V_M)$; the (random) resample size M^* is a realization of a Poisson distribution with parameter M. We note in passing that $\hat{\lambda}$ may not exist in exceptional cases. However, in the simulations reported in Section 4 $\hat{\lambda}$ could always be determined. The numerical grid of λ-values was taken to be equally spaced w.r.t. the corresponding u_λ-values with (constant) width .01. This amounts to differences between subsequent λ-values not bigger than 2x10^{-4} in our simulations. A minor difficulty arises when $M^* = 0$, i.e. there is no bootstrap sample and we simply delete such 'empty' resamples; accordingly the P_n^*-probability (3.1) is estimated in such cases by the number of times the inequality in (3.1) is valid divided by the number of 'non-empty' bootstrap samples. Note that, when $M \geq 5$, the probability that $M^* = 0$ is at most equal to $e^{-5} = 0.0067$.

Obviously $\hat{\lambda}$ will typically be somewhat smaller then α, and the calibrated confidence bound

$$\hat{D}_n + \hat{c}_{2,n,\hat{\lambda}}\frac{N}{n}(\sum_{j=1}^{M} V_j^{2})^{\frac{1}{2}} \tag{3.2}$$

will usually be larger than (2.11), but the calibrated upper bound (3.2) possesses the benificial property of having estimated confidence level at least equal to $1-\alpha$. Our bootstrap estimate (3.1), with $\lambda = \alpha$, may be used as a diagnostic tool to check whether the empirical Cornish-Fisher bound has already the desired confidence level $\geq 1-\alpha$, and calibration of the bound (2.11) would not be needed. In contrast to (2.11), the bound (3.2) requires a lot of computation, as it involves extensive bootstrapping. In practice, however, bootstrap calibration will only be needed, when the real data set at hand contains not too many errors and/or the observed non-zero error amounts in the sample contain one or more extreme 'outliers'. Otherwise , it is to be expected that the computationally much simpler bound (2.11) will usually suffice. One may try to develop a practical guideline for the use of bootstrap calibration in our setting (cf. also Young (1994) p.411, for similar advice). In any case, the computationally very demanding 'double bootstrap' technique is avoided as our starting interval (2.11) is a non-bootstrap interval. For this very reason we haven't used the studentized bootstrap (cf Hall (1988), Helmers (1991)) in the first place, but instead relied on an empirical Cornish-Fisher expansion.

In the case $M = 0$ the confidence upper bound (2.11) takes the value infinity, as is to be expected, because when $M = 0$ nothing can be said about an upper bound for D with confidence. In fact, if $M = 0$ (as often occurs in auditing practice) one has to impose a restriction on the error structure to make progress.

4. Simulations

In this section we briefly describe some Monte Carlo simulations which were carried out to assess the performance of our new upper confidence limit for various audit populations of practical interest. The size of the finite populations under consideration was set to be equal to $N = 5\text{x}10^5$.

In our first simulation we take p=0.02, n=500 and $G = exp(200)$, the exponential distribution with mean 200. The errors e_i are distributed according to the non-standard mixture distribution

$$F = 0.02\ exp(200) + 0.98\delta_0. \tag{4.1}$$

This first example represents a relatively simple audit population. The parameter of interest D is

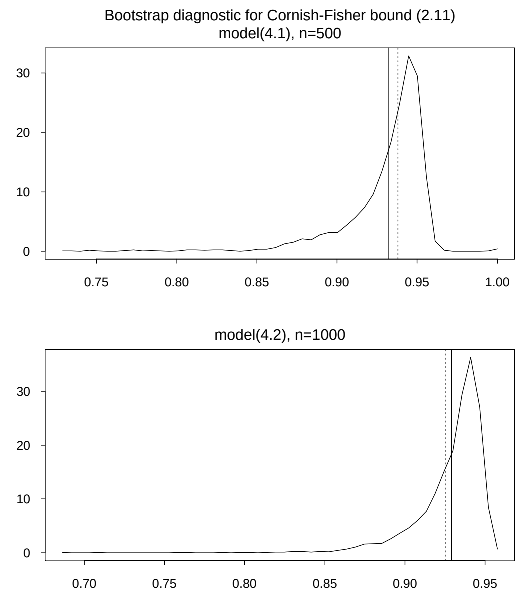

Figure 1: Density of 2000 bootstrap estimates (3.1), with $\lambda = \alpha$, of true coverage probability; true coverage probability (dashed line) and average of 2000 bootstrap estimates (solid line)

replaced by $E_F D = \frac{N}{n}\nu\mu$, which equals 2x$10^6$. Note that the number of non-zero errors in our sample of size 500 from F is Poisson distributed with mean 10. The true (nominal) coverage probability (3.1), with $(1-\alpha)$=0.95, was estimated accurately by Monte Carlo, using 5x10^5 samples from F, to be 0.938. Next, on the basis of a single sample of size 500 from F, the bootstrap estimate (3.1), with $\lambda = \alpha$, of the coverage probability was computed, using $B = 5000$ bootstrap resamples, with random resample size Poisson(M), where M denotes the number of errors in the original sample from F. This procedure was repeated 2000 times. The average of these 2000 bootstrap estimates of the *true* coverage probability 0.938 equals 0.932, while a density plot of these estimates is given in Figure 1 (upper panel). It can be seen from the graph that our bootstrap estimate for the coverage probability of the Cornish-Fisher bound reflects about 84% of the time the fact that our upper confidence limit (2.11) has a true confidence level somewhat smaller than 0.95, namely 0.938. Hence, computing (3.1), with $\lambda = \alpha$, yields a fairly reliable diagnostic for the validity of the Cornish-Fisher upper bound (2.11) in this case. Calibration is perhaps needed here, as 0.932 falls short of 0.95.
In a second simulation we consider a more realistic non-standard mixture:

$$F = 0.02\ exp\left(\frac{100}{3}\right) + 0.01\ exp\left(\frac{1000}{3}\right) + 0.97\delta_0. \tag{4.2}$$

In this setup we take into account the possibility of 'outliers' among the observed non-zero error amounts, by assuming that G consists of a mixture of two exponentials, with means $\frac{100}{3}$ and $\frac{1000}{3}$ respectively. In the present example we take $n = 1000$ and $E_F D = \frac{N}{n}(\nu_1\mu_1 + \nu_2\mu_2)$, with $\nu_1 = 20$, $\mu_1 = \frac{100}{3}$, $\nu_2 = 10$, $\mu_2 = \frac{1000}{3}$, which is again equal to 2x10^6. The number of non-zero errors in our sample of size 1000 from F is now Poisson distributed with mean 30; on the average 10 of these will be 'outliers'. The true (nominal) coverage probability, with $(1-\alpha)$=0.95, was estimated by

Monte Carlo, using 5x10^5 samples from F, to be 0.925. Next (3.1), with $\lambda = \alpha$, was estimated 2000 times, employing 2000 samples of size 1000 from (4.2) and using $B = 5000$ bootstrap resamples each time. The results are summarized in Figure 1 (lower panel). The bootstrap diagnostic works well.

Acknowledgements
This paper was written as part of a research project with the Statistical Audit Group of PricewaterhouseCoopers, Amsterdam. I thank Dick van der Hoeven for suggesting the problem. Comments of Harmen Ettema and John Haworth are gratefully acknowledged. Rob van der Horst (CWI) carried out the simulations. I thank the Editor and two referees for their very valuable remarks.

References

A.D.Barbour, L.Holst, S.Janson (1992), *Poisson Approximation*, Oxford Studies in Probability, Clarendon Press, Oxford.

R.Beran (1987), Prepivoting to reduce level error of confidence sets. Biometrika 74, 457-468.

P.Hall (1987), Edgeworth expansion for Student's t statistic under minimal moment conditions, *Ann.Probab.* 15, 920-931.

P.Hall (1988), Theoretical comparison of bootstrap confidence intervals (with discussion), *Ann. Statist.* 16, 927-985.

P.Hall and M.A.Martin (1988), On bootstrap resampling and iteration. Biometrika 75, 661-671.

R.Helmers (1991), On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistic, *Ann.Statist.*, 19, 470-484.

J.L.Jensen (1995), *Saddle point approximations.* Oxford Science Publications.

H.Tamura (1989), Statistical Models and Analysis in Auditing, Panel on Nonstandard Mixtures of Distributions, *Statistical Science* 4, 2-33.

G.A.Young (1994), Bootstrap: more than a stab in the dark? (with Discussion) *Statistical Science* 9, no 3, 382–415.